말소리와 음성과학 제7권 제4호 (2015.12.31)
ISSN 2005-8063, pp. 27~33
ksss.jams.or.kr
http://dx.doi.org/10.13064/KSSS.2015.7.4.027

깊은 신경망을 이용한 오디오 이벤트 분류

Audio Event Classification Using Deep Neural Networks

임 민 규[1] · 이 동 현[2] · 김 광 호[3] · 김 지 환[4]
Lim, Minkyu · Lee, Donghyun · Kim, Kwang-Ho · Kim, Ji-Hwan

ABSTRACT

This paper proposes an audio event classification method using Deep Neural Networks (DNN). The proposed method applies Feed Forward Neural Network (FFNN) to generate event probabilities of ten audio events (dog barks, engine idling, and so on) for each frame. For each frame, mel scale filter bank features of its consecutive frames are used as the input vector of the FFNN. These event probabilities are accumulated for the events and the classification result is determined as the event with the highest accumulated probability. For the same dataset, the best accuracy of previous studies was reported as about 70% when the Support Vector Machine (SVM) was applied. The best accuracy of the proposed method achieves as 79.23% for the UrbanSound8K dataset when 80 mel scale filter bank features each from 7 consecutive frames (in total 560) were implemented as the input vector for the FFNN with two hidden layers and 2,000 neurons per hidden layer. In this configuration, the rectified linear unit was suggested as its activation function.

Keywords: audio event classification, deep neural networks, mel scale filter bank

1. 서론

스마트폰의 다양성, 다목적성, 휴대성으로 개인의 동영상 촬영이 일상의 한 부분이 되면서 우리가 다루는 정보가 텍스트에서 영상으로, 대중미디어에서 개인미디어로 진화하고 있으나, 아직까지는 미디어 분류의 대부분을 메타데이터에 의존하고 있다. 이에 맞추어 구글, 페이스북 등 영상을 이용한 새로운 맞춤지식을 생성하는 기술 개발이 시도되는 등 최근 미디어 분석 연구가 활발히 진행되고 있다. 영상속의 의미를 자동으로 분석하기 위해서는 그 속에 포함된 오디오 이벤트를 인식하는 기술은 필수적이다. 오디오 이벤트 인식의 경우 기존에는 오디오 신호로부터 zero crossing rate, spectal flux, band periodicity 등 다양한 특징 값들의 성능을 검증하는 연구와, 전통적인 분류 방법인 규칙기반 (rule-based), Gaussian Mixture Model (GMM) 기반 분류기에 관련한 연구가 주를 이루었다[1]-[3]. 하지만 대부분의 연구는 음악/음성/기타소리를 구분하는 등 제한적인 클래스 분류가 주를 이루었다.

최근 기계학습 분야에서 괄목할만한 성능 향상을 보이는 기술로서 Deep Neural Network (DNN)이 주목 받고 있다. DNN은 많은 수의 계층으로 구성된 깊은 인공 신경망으로서 기존의 인공 신경망보다 복잡한 비선형적인 학습 경계를 구분 지을 수 있어 분류 문제에 있어 더 좋은 성능을 얻을 수 있다. 다만 DNN의 수많은 파라미터를 추정하는 데에 있어서 높은 연산량이 요구되어 어려움이 있었지만, 최근 하드웨어 기술의 발전으로 다양한 응용 분야에 DNN을 성공적으로 적용할 수 있게 되었다.

DNN은 음성인식 및 이미지 분류에 적용되어 많은 성능향상을 보였으나, 오디오 이벤트 분류에 적용된 사례는 많지 않다. 본 논문에서는 DNN을 이용한 오디오 이벤트 분류기를 구현하고, DNN을 구성하는 하이퍼파라미터를 실험적으로 추정한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 오디오 이벤트 인식을 위한 기존 연구들에 대하여 서술하고, 3장에서는 DNN을 이용한 오디오 이벤트 분류기에 대해 서술한다. 4

1) 서강대학교, lmkhi@sogang.ac.kr
2) 서강대학교, redizard@sogang.ac.kr
3) 서강대학교, kimkwangho@sogang.ac.kr
4) 서강대학교, kimjihwan@sogang.ac.kr, 교신저자

접수일자: 2015년 11월 21일
수정일자: 2015년 12월 06일
게재결정: 2015년 12월 17일

장에서는 DNN을 이용한 오디오 이벤트 분류기의 하이퍼파라미터 파라미터를 실험적으로 추정하고 기존 분류기와의 비교를 통해 성능향상의 정도를 평가하고자 한다. 5장에서는 결론을 맺는다.

2. 기존 연구

HMM-SVM 기반의 오디오 이벤트 분류기에 대한 연구에서는 15개의 오디오 이벤트 분류기를 Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP), Zero Crossing Rate (ZCR) 등 다양한 특징벡터를 조합하여 학습시켰으며, 특징벡터의 선택에 따라 인식 성능의 차이가 있음을 보였다[4]. 영화, 다큐멘터리, 토크쇼, 뉴스에서 등장하는 오디오 이벤트들을 직접 레이블링하여 이벤트들의 검출율을 측정하였다. 실험 결과 PLP를 특징벡터로 사용한 경우 성능이 가장 높았다.

하나의 샘플에 여러 개의 이벤트가 있는 경우에 대한 오디오 이벤트 시퀀스 분류 연구가 있었다[5]. 기존의 오디오 이벤트 분류의 경우 하나의 오디오 샘플에는 하나의 이벤트만 존재하는 것으로 가정되었지만 실제로는 여러 이벤트가 나열되는 경우가 많기에 이벤트 시퀀스에 대한 분류를 시도하였다. 각 오디오 이벤트들을 GMM을 통하여 모델링을 한 후, 이벤트 시퀀스 분류를 위하여 3-state 기반의 Hidden Markov Model (HMM)을 사용하였다. 이는 오디오 이벤트가 일련의 순서로 등장할 때 성능이 좋은 장점이 있는 반면 다양한 오디오 이벤트들이 레이블링 되어있는 학습 자료를 수집하기 어려운 단점이 있다.

Deep Belief Network (DBN)을 음악 장르 구분 및 음악가 분류 문제에 적용한 연구가 있었다[6]. DBN은 레이블링 되지 않은 다량의 학습자료를 이용하여 은닉층을 학습시킨 후에, 소량의 레이블링된 자료를 이용하여 출력층을 학습시키는 방법이다. DBN의 은닉층 학습은 Restricted Boltzmann Machine (RBM)을 이용하여 무감독 학습시킨 은닉층을 greedy 방법으로 층층이 쌓는다. 최종적으로 출력층에 대해서에만 소량의 레이블링된 학습 자료를 이용하여 감독 학습시킨다. 실험 결과 5개의 음악 장르 구분에 대하여 약 73%의 분류 성능을 보였고, 음악가 분류에 적용 시 4개의 음악가 분류문제에서는 약 80%의 분류성능을 보였다. 이 연구는 레이블링 되지 않은 다량의 학습자료를 이용할 수 있는 장점을 보였지만, 분류 실험의 클래스 수가 작다는 한계가 있다.

RBM을 이용한 오디오 또 다른 이벤트 분류 연구로서, Crowd, Traffic, Applause, Music 네 개의 이벤트에 대해서 구분짓는 연구가 있었다[7]. 이 연구에서는 은닉층의 경우 입력벡터에 대한 출력벡터를 생성하도록 RBM을 이용하여 학습시킨 후, 최상위 층에 대해서만 FFNN을 적용하여 오디오 이벤트를 출력하도록 하였다. RBM을 이용한 DNN을 GMM 및 SVM과 비교 평가를 하였고, RBM의 분류 성능이 SVM, GMM 보다 더 높게 나왔다. 하지만 역시 오디오 이벤트의 수가 적다는 문제가 있다.

축구 중계 영상에서 오디오 정보를 이용하여 다섯 개의 이벤트 분류에 DBN을 사용한 연구가 있다[8]. 다섯 개의 이벤트는 해설, 관중소리, 해설+관중소리, 흥분된 해설이며, DBN을 학습시켜 SVM 분류기와의 성능을 비교 평가하였다. 실험 결과 SVM의 성능이 DNN보다 약간 더 높았으며, 적은 양의 학습자료로 인한 결과로 분석되었다.

기존의 SVM과 DNN의 차이를 이론적으로 분석한 연구가 있었다[9]. 이 연구에서는 SVM과 DNN을 shallow architecture와 deep architecture로 구분 짓는다. SVM의 경우 커널 함수를 통하여 차원을 줄이며 클래스 분포를 구분짓는 선을 긋는 방식이며 이것은 하나의 은닉층을 가지는 인공신경망의 한 형태로 볼 수 있다. 반면 여러개의 은닉층을 가지는 DNN의 경우 비선형적인 경계를 계층적으로 쌓기 때문에 SVM 혹은 단일 은닉층의 인공신경망보다 더 복잡한 decision boundary를 구분지을 수 있다. 한 예로, 단일 은닉층의 인공신경망으로는 XOR 구분 문제를 모델링 할 수 없지만, 여러 은닉층의 인공신경망으로는 구분할 수 있다. 다만 DNN의 파라미터들을 학습하기 위해서는 다량의 자료가 필요하다.

오디오 정보 분석에 관련한 과제들은 대부분 음성, 음악에 대한 내용이 주류이다. CHiME는 잡음환경 하에서의 키워드 인식이 주된 내용이다[10]. 다양한 잡음 환경에서의 키워드 인식 성능을 사람의 인식 성능 측정치와 비교 평가하였고, 가장 높은 성능의 인식기의 경우 사람대비 5%정도의 인식률 차이가 있었다. MIREX는 음악의 비트 트래킹, 화음 추정, 멜로디 추출, 장르 구분 등 MIR (Music Information Retrieval)에 관련한 태스크를 수행한다[11]. CLEAR 과제는 비디오 / 오디오의 멀티 모달 정보로부터 사람의 행동, 반응, 주변 환경 정보등을 인식하는 시스템을 평가하는 과제이다. 여기에는 오디오 환경 분석이 포함되어 있고, CMU에서는 공항/길거리/버스 등 9개의 환경에 대하여 소리를 통해 인식하는 HMM기반 시스템을 평가 하였고, 약 15%수준의 오류율을 보였다[12]. TRECVID 과제에서는 주로 비디오에 대한 분석이 이루어지고 있으며, Semantic Indexing (SIN), Multimedia Event Detection (MED), Localization (LOC) 등의 세부 과제가 있다[13]. MED의 경우 오디오 이벤트 분석이 포함되어 있고, SiSEC 과제는 음악과 음성이 혼합된 오디오에서의 소스 구분문제에 초점을 두고 있다[14].

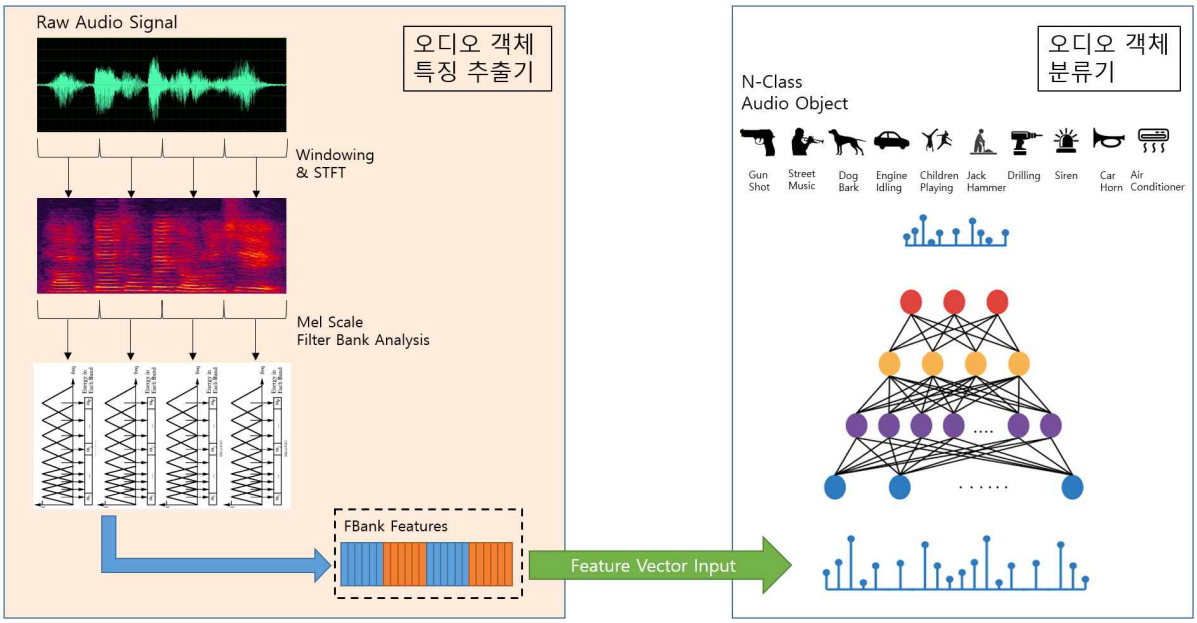

그림 1. DNN을 이용한 오디오 이벤트 분류기 구조
Figure 1. Structure of audio object classifier using DNN

3. DNN 기반 오디오 이벤트 분류기

제안한 DNN 기반 오디오 이벤트 분류기는 두 가지 모듈로 구성되어 있다. 하나는 오디오 특징 추출기이며 다른 하나는 오디오 이벤트 분류기이다. <그림 1>은 DNN 기반 오디오 이벤트 분류기의 구조를 보여준다.

3.1 오디오 특징 추출기

오디오 특징 추출기는 DNN의 입력벡터에 사용되는 벡터열을 생성한다. 입력된 오디오 (2-byte per sample, mono, 16kHz)에 대하여 20ms 길이의 해밍 (Hamming) 윈도우가 10ms 단위로 이동하면서 Short Time Fourier Transform (STFT) 이 수행된다. 매 윈도우마다 mel scale로 증가하는 삼각 형태의 bin을 씌워 각 주파수의 에너지마다 가중치를 곱하여 특징 값을 추출한다. 이를 하나의 벡터로 표현하여 Mel-scale Filter Bank (FBANK) 특징벡터를 생성한다.

하나의 윈도우로부터 추출된 특징벡터는 10 ms 에 해당되는 특징 값으로서 이는 소리에 대한 매우 일시적 특징만을 가지며 동일한 소리의 범위 내에서 윈도우 이동에 따른 특징 변화가 매우 크다. 따라서 컨텍스트 정보를 반영하기 위하여 현재 윈도우 기준 좌/우 N개의 윈도우로부터 추출된 특징벡터를 결합하여 하나의 벡터를 생성하며 이를 분류기의 입력으로 사용한다. 고려하는 윈도우 크기에 따라 DNN 기반 오디오 이벤트 분류기의 입력되는 입력벡터 크기가 달라진다. 이의 변화에 따른 성능평가는 4장에서 기술한다.

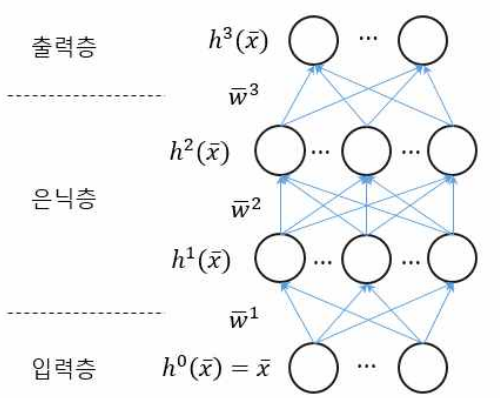

그림 2. 은닉층이 2개인 FFNN의 예
Figure 2. Example of 2-Hidden layer FFNN

3.2 오디오 이벤트 분류기

본 절에서는 오디오 이벤트 분류기의 구조를 서술한다. 오디오 이벤트 분류기는 <그림 1>에서와 같이 Feed Forward Neural Network (FFNN)으로 이루어져 있다. FFNN은 하나씩의 입력층과 출력층, 그리고 하나 이상의 은닉층으로 구성되어 있다. 각각의 층은 뉴론들로 구성되어 있으며, 하나의 뉴론은 weight와 bias를 파라미터로 가지고, 아래층의 모든 뉴론들의 출력이 해당 뉴론의 입력이 된다. 3.1절의 오디오 특징 추출기로부터 생성된 특징벡터는 오디오 이벤트 분류기를 구성하는 FFNN의 입력벡터 $\bar{x}$가 된다. 입력벡터 $\bar{x}$로부터 수식 (1)과 수식 (2)를 통하여 FFNN의 최종 출력층까지 순전파 된다. 입력벡터에 대한 출력층에서의 값은 각 클래스에 대한 발생확률을

의미한다. <그림 2>은 은닉층 수가 2인 FFNN의 예를 보여 준다. <표 1>은 수식에서 사용된 기호의 의미를 정의한다.

$$a_i^k(\overline{x}) = b_i^k + \sum_{j=1}^{N^{k-1}} w_j^k h_j^{k-1}(\overline{x}) \; (1 \le k \le L+1) \tag{1}$$

$$h_i^k(\overline{x}) = g(a_i^k(\overline{x}))(1 \le k \le L+1) \tag{2}$$

$$\begin{aligned} P(y=i|\overline{x}, W,b) &= h_i^{L+1}(\overline{x}) \\ &= softmax(a_i^{L+1}(\overline{x})) \\ &= \frac{e^{a_i^{L+1}(\overline{x})}}{\sum_{j=1}^{N^{L+1}} e^{a_j^{L+1}(\overline{x})}} \end{aligned} \tag{3}$$

표 1. FFNN 수식 관련 기호 정의
Table 1. Notations in FFNN equations

수식 기호	정의
$\overline{x}$	입력 벡터
L	은닉층 수
$\overline{w}^k$	*k*-번째 층, 모든 뉴론의 weight 벡터
w_i^k	*k*-번째 층, *i*-번째 뉴론의 weight 값
W	FFNN의 모든 weight 집합
b	bias
N^k	*k*-번째 층의 뉴론 수
$a()$	preactivation 함수
$g()$	활성함수
$h_i^k(\overline{x})$	입력 벡터 $\overline{x}$에 대한 *k*-번째 층, *i*-번째 뉴론의 출력 값
y	분류기의 출력 클래스

수식 (1)에서 $a_i^k(\overline{x})$는 입력벡터 $\overline{x}$가 들어왔을 때 k 번째 층의 i 번째 뉴론에서의 선활성함수 계산 결과이다. 해당 뉴론의 최종 출력값은 선활성함수의 결과에 활성함수를 취한 결과가 된다. 은닉층의 활성함수로는 주로 tanh, sigmoid 등의 비선형 함수를 사용한다. 비선형 활성함수는 선형 활성함수보다 복잡한 분류 경계를 표현할 수 있기 때문에 모델 성능이 높은 것으로 알려져 있다[15]. 최근에는 Rectified Linear Unit (ReLU)가 비선형 활성함수로써 주목받고 있다[16]. ReLU 함수의 수식은 $g(x) = \max(0,x)$와 같으며 tanh, sigmoid 등과 비교하여 연산이 간단하며 은닉층이 많은 경우 학습 시 발생하는 vanishing gradient 문제를 감소시키는 이점이 있다. 출력층의 활성함수는 softmax 가 사용되며 최종 출력값은 수식 (3)에 의해 계산된다. 모델 파라미터 W,b 와 입력벡터 $\overline{x}$가 주어졌을 때 각 클래스에 대한 발생확률이 출력된다. 분류기는 입력벡터에 대하여 가장 큰 발생확률을 가지는 클래스를 분류 결과로 제시한다.

파라미터 학습은 수식 (5)와 같은 Negative Log Likelihood (NLL)를 손실 함수로 정의하여 모든 학습자료에 대하여 수식 (4)의 손실 값이 최소가 되도록 한다. 학습 알고리즘은 stochastic gradient descent 방법을 취한다[17].

$$E(\Theta, D) = NLL(\Theta, D) + \lambda \| \Theta \|_p^p \tag{4}$$

$$NLL(\Theta, D) = -\sum_{s=1}^{|D|} \log P(y^{(s)}|\overline{x}^{(s)}, \Theta) \tag{5}$$

$$\| \Theta \|_p = \left(\sum_{j=1}^{|\Theta|} |\Theta_j|^p \right)^{\frac{1}{p}} \tag{6}$$

위의 수식에서 D는 모든 학습 자료로부터 추출된 특징벡터 열 집합를 의미하고, Θ는 앞서 학습된 모든 파라미터 (weight 및 bias) 집합을 의미한다. j는 모든 파라미터의 index정보를 표현한다. 수식 (6)에서의 $\| \Theta \|_p$는 학습자료에 대한 과적합 (overfitting)을 방지하기 위한 regularizer 이다.

테스트 샘플에 대한 이벤트 분류 결과는, 해당 샘플의 모든 프레임에 대한 분류 결과로부터 각 이벤트들의 발생 확률를 모두 누적한 결과가 가장 큰 발생 확률 값을 가지는 이벤트가 최종 출력 된다.

4. 하이퍼파라미터 변화에 따른 오디오 이벤트 분류 성능 평가

음성 및 이미지 인식의 경우 Linguistic Data Consortium (LDC), 이미지넷 등 학술적으로 비교 평가할 수 있는 공통의 학습자료가 존재하는 반면, 오디오 이벤트 분류의 경우 대부분의 연구가 별도로 자료를 수집하여 검증하기 때문에 공통된 평가 지표를 구하기 어렵다. 본 연구에서는 접근 가능한 자료 중 최근 연구로서 평가 지표가 존재하는 학습자료 중 하나를 선택하였으며[18], 해당 연구의 SVM 기반 오디오 이벤트 분류기의 성능은 약 70%를 나타내었다. 본 연구에서는 SVM 기반 분류기를 베이스라인으로 삼는다.

표 2. UrbanSound8K dataset의 이벤트 종류
Table 2. Event list of UrbanSound8K dataset

Air conditioner	Engine idling
Car horn	Gun shot
Children playing	Jack hammer
Dog barks	Siren
Drilling	Street music

학습 및 평가에 사용한 자료는 UrbanSound8K 이다[18]. UrbanSound8K는 www.freesound.org로부터 수집한 오디오 신호를 이벤트 구간에 맞추어 labeling한 자료 이다. 총 8,732의 샘플로 구성되어 있으며 하나의 샘플은 4초 이하로 제공된다. 오디오 이벤트 종류는 총 10 개 이며 약 9시간 분량의 자료로 구성되어 있다. <표 2>는 UrbanSound8K에서 구성된 오디오 이벤트 종류를 보여준다. 모든 자료는 16bit-mono의 16kHz로 일괄 변환하였고 자료의 9/10은 학습에 사용되었고, 1/10은 평가에 사용되었다.

표 3. 하이퍼파라미터 변화에 따른 오디오 이벤트 분류 성능
Table 3 Audio event classification accuracy according to various hyper parameter

은닉층 수	은닉층 당 뉴론 수	활성 함수	FBank 차수	입력 프레임 개수 (2N+1)	인식률 (%)
2	500	Tanh	40	15	54.6
2	500	ReLU	40	15	56.0
2	1,000	ReLU	40	15	69.4
2	2,000	ReLU	40	15	75.9
2	2,000	ReLU	80	7	79.2
3	1,500	ReLU	40	15	75.1

특징 벡터 추출은 음성인식 툴킷인 HTK[19] 를 사용하였고, DNN기반 오디오 이벤트 분류기는 Theano[20] 를 사용하여 구현하였다. 학습 시 minibatch 크기는 98,569 이며, 초기 learning rate은 0.01로 시작하였고, validation error가 30 회 동안 감소하지 않으면 learning rate을 10%씩 감소시켰다. 최소 learning rate 은 0.001 이다. 수식 (4), (6)에서 $\lambda = 0.001$, $p = 2$ 로 설정 하였다.

DNN 모델의 하이퍼파라미터인 은닉층 당 뉴론 수, 은닉층 수, 활성함수 종류, 입력윈도우 숫자 등을 다르게 적용하여 실험을 진행 하였다. 그에 대한 실험 결과는 <표 3>과 같다.

실험 결과 은닉층의 활성함수는 tanh 함수보다 ReLU 함수를 사용하는 것이 전반적으로 높게 측정 되었고 이 후로는 은닉층의 활성함수를 ReLU를 사용한 모델에 대하여 분류 정확도를 평가하였다.

2-은닉층 / 2,000-은닉층 당 뉴론 수 / ReLU 활성함수를 사용하여 학습한 모델이 가장 높은 성능을 보였고, 특징벡터로는 40 차보다 80차를 사용한 경우의 성능이 더 높게 측정되었다. 위의 모델보다 모델 크기를 늘렸을 경우 (3-은닉층 / 1,500-은닉층 당 뉴론 수) 분류 성능이 오히려 감소하는 것을 확인 할 수 있다. 이는 현재 사용하고 있는 학습 자료량으로서는 인식 성능이 모델 크기 대비 수렴했다고 볼 수 있다.

<표 4>는 가장 높은 인식률을 보인 하이퍼파라미터 오디오 이벤트 분류 결과에 대한 confusion matrix를 보여준다. Confusion matrix 상의 값은 인식률을 가장 가까운 정수로 매핑한 결과이다. 결과를 보면 dog barks, drilling, engine idling에 대한 결과로 air conditioner가 도출 되는 경우가 많이 있는데, engine idling과 drilling의 경우는 소리가 유사한 경우가 많았다. 하지만 dog barks의 경우는 소리 자체는 전혀 다르지만, 하나의 샘플 상에서 개소리가 포함되어 있는 부분이 무음 구간보다 적은 경우가 많이 있었다. 이런 경우 무음구간에 대한 결과가 air conditioner로 많이 매핑되는 경우가 발생하였다. 따라서 무음 구간에 대한 처리가 필요할 것으로 보인다.

동일한 학습 자료를 사용한 SVM 분류기[18]의 경우 70%의 최대 성능을 보인 점을 감안하면 제안하는 DNN기반 오디오 이벤트 분류기는 약 10%정도의 분류성능 개선이 있었다. 반면

표 4. 2-은닉층 / 2,000-은닉층 당 뉴론 수 / ReLU 활성함수로 학습한 모델의 오디오 이벤트 분류 결과에 따른 confusion matrix
Table 4. Confusion matrix for audio event classification with 2 hidden layer / 2,000 neurons per layer / ReLU function

	air conditioner	car horn	children playing	dog barks	drilling	engine idling	gun shot	jack hammer	siren	street music
air conditioner	**88**	0	0	0	3	9	0	0	0	0
car horn	3	**64**	6	0	12	6	0	0	0	9
children playing	9	0	**72**	3	2	1	1	0	0	12
dog barks	12	0	7	**69**	0	4	0	0	2	6
drilling	12	0	1	4	**74**	7	0	0	1	1
engine idling	16	0	0	0	0	**82**	0	1	0	1
gun shot	3	0	0	13	0	0	**84**	0	0	0
jack hammer	0	0	0	0	1	14	0	**80**	0	5
siren	4	0	1	4	0	0	0	0	**87**	4
street music	0	0	9	0	0	0	0	0	1	**90**

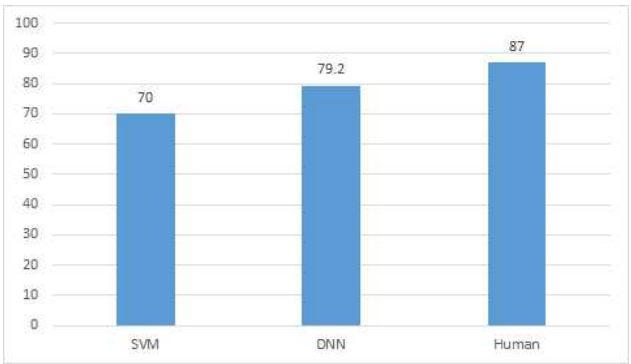

그림 3. SVM기반 분류기 / DNN기반 분류기 / 사람의 오디오 이벤트 분류 성능
Figure 3. Accuracy of SVM classifier / DNN classifier / Human

동일한 테스트자료에 대한 두 명의 오디오 이벤트 전문가의 분류 정확도를 직접 테스트 하여 측정해 본 결과 약 87%를 보였고, 이를 통해 아직 사람의 분류 수준에는 미치지 못하는 것으로 보인다. <그림 3>은 동일 자료를 통하여 평가한 SVM 기반 분류기와 제안하는 DNN 기반 분류기, 그리고 사람이 직접 평가한 성능을 보여준다.

5. 결론

본 연구에서는 DNN을 이용하여 오디오 이벤트 분류기를 제안하였고, 실험을 통하여 그 성능을 확인하였다. 다양한 하이퍼파라미터 상에서 학습한 DNN 기반 분류기를 통해 성능을 측정하였다. 실험 결과 79.2%의 최대 성능을 보였으며 이는 동일한 코퍼스를 사용한 SVM 기반 분류기 보다 13.1%의 상대적

성능향상을 보였다. 추후에는 입력 샘플의 무음구간에서 발생하는 오류를 보정하기 위한 처리에 대한 연구와 더 다양한 하이퍼파라미터 조합에 대한 추가적인 실험을 진행할 계획이다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 일환으로 하였음. [R0126-15-1112, 퍼스널 미디어가 연결공유결합하여 재구성 가능케 하는 복합 모달리티 기반 미디어 응용 프레임워크 개발]

참고문헌

[1] Lu, L., Jiang, H., & Zhang, H. (2001). A robust audio classification and segmentation method, in *Proc. ACM International Conference on Multimedia*, 203-211.

[2] Xu, M., et al. (2003). Creating audio keywords for event detection in soccer video, in *Proc. IEEE International Conference on Multimedia and Expo*, 281-284.

[3] Cheng, W., Chu, W., and Wu, J. (2003). Semantic context detection based on hierarchical audio models, in *Proc. ACM SIGMM International Workshop on Multimedia Information Retrieval*, 109-115.

[4] Elo, J. P., et al. (2009). Non-speech audio event detection, in *Proc. Internationa Conference on Acoustics, Speech and Signal Processing*, 1973-1976.

[5] Heittola, T., et al. (2013). Context-dependent sound event detection, *EURASIP Journal on Audio, Speech, and Music Processing*, 1 1-13.

[6] Lee, H., Pham, P., Largman, Y., & Ng, A. Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. in *Proc. Advances in Neural Information Processing Systems,* 1096-1104.

[7] K, Zvi., & T, Orith. (2013). Audio event classification using deep neural networks, in *Proc. INTERSPEECH*, 1482-1486.

[8] Ballan, L., et al. (2009). Deep networks for audio event classification in soccer videos, in *Proc. International Conference on Multimedia and Expo*, 474-477.

[9] Bengio, Y. & LeCun, Y. (2007). Scaling learning algorithms towards AI, *Large-scale Kernel Machines*, Vol. 34, No.5, 321-360.

[10] Barker, J., et al. (2012). The PASCAL CHiME speech separation and recognition challenge, *Computer Speech & Language*, Vol. 27, No. 3, 621-633.

[11] Downie, S., et al. (2010). The Music Information Retrieval Evaluation eXchange: Some observations and insights, *Advances in Music Information Retrieval.* Springer, 93-115.

[12] Malkin, R. G. (2007). *Multimodal Technologies for Perception of Humans.* Springer, 323-330.

[13] Smeaton, F. et al. (2006). Evaluation campaigns and TRECVid, in *Proc. ACM International Workshop on Multimedia Information Retrieval*, 321-330.

[14] Vincen, E., et al. (2012). The signal separation evaluation campaign (2007‑2010): Achievements and remaining challenges, *Signal Processing*, Vol. 92, No. 8, 1928-1936.

[15] Larochelle, H., et al. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. in *Proc. International Conference on Machine learning,* 473-480.

[16] Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout, in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 8609-8613.

[17] Bottou, L. (2004). *Advanced Lectures on Machine Learning*, Sringer, 146-168.

[18] Salamon, J., Jacoby, C., & Bello, J. P. (2014), A dataset and taxonomy for urban sound research, in *Proc. ACM International Conference on Multimedia*, 1041-1044.

[19] Young, S., et al. (1999). *The HTK Book.* Cambridge, U.K.: Entropic.

[20] Bergstra, J., et al. (2010). Theano: A CPU and GPU math expression compiler. in *Proc. Python for Scientific Computing Conference*, Vol. 4, p. 3.

• **임민규 (Lim, Minkyu)**
서강대학교 컴퓨터공학과
서울시 마포구 백범로 35(신수동)
Tel: 02-715-2715 Fax: 02-704-8273
Email: lmkhi@sogang.ac.kr
관심분야: 음성인식, 오디오 이벤트 분류
현재 컴퓨터공학과 대학원 박사과정 재학 중

• **이동현 (Lee, Donghyun)**
서강대학교 컴퓨터공학과
서울시 마포구 백범로 35(신수동)
Tel: 02-715-2715 Fax: 02-704-8273
Email: redizard@sogang.ac.kr
관심분야: 음향모델, 대화모델
현재 컴퓨터공학과 대학원 석박사 통합 과정 재학 중

• **김광호 (Kim, Kwang-Ho)**
서강대학교 컴퓨터공학과
서울시 마포구 백범로 35(신수동)
Tel: 02-715-2715 Fax: 02-704-8273
Email: kimkwangho@sogang.ac.kr

관심분야: 음성인식, 언어모델
현재 컴퓨터공학과 대학원 박사과정 재학 중

- **김지환 (Kim, Ji-Hwan), 교신저자**
서강대학교 컴퓨터공학과
서울시 마포구 백범로 35(신수동)
Tel: 02-705-8924 Fax: 02-704-8273
Email: kimjihwan@sogang.ac.kr
관심분야: 음성인식
현재 컴퓨터공학과 부교수 재직 중